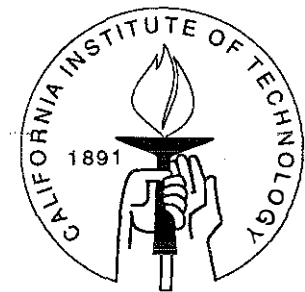


DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES
CALIFORNIA INSTITUTE OF TECHNOLOGY
PASADENA, CALIFORNIA 91125

THE EXTRACTION OF INFORMATION FROM MULTIPLE POINT ESTIMATES

Mahmoud A. El-Gamal



SOCIAL SCIENCE WORKING PAPER 839

February 1993

The Extraction of Information from Multiple Point Estimates

Mahmoud A. El-Gamal

Abstract

The use of a number of point estimation experiments to construct a least informative prior subject to the information in the estimation experiments is studied. The form of the resulting prior is established. The prior depends on parameters that result from solving a calculus of variations problem. It is shown that simple Gibbs sampler algorithms converge to the desired solution, and simulated annealing algorithms yield the mode of the prior. The Gibbs sampler algorithm is ergodic, and hence Bayes risks can be directly computed using time averages of a single series of draws from the sampler.

Keywords: Information theory, meta-analysis, Gibbs sampler, simulated annealing.

Forthcoming: Journal of Nonparametric Statistics.

The Extraction of Information from Multiple Point Estimates

Mahmoud A. El-Gamal

1 Introduction

In this paper, we consider the following problem. A number of point estimation experiments have been conducted to achieve point estimates of a particular parameter $\theta \in \Theta$. We wish to find a Bayesian prior on the parameter space $(\Theta, \mathfrak{B}(\Theta))$ (where $\Theta \subset \mathbb{R}^l$ and $\mathfrak{B}(\Theta)$ is the Borel σ -algebra of subsets of Θ) based on the information contained in the outcomes of the point estimation experiments. We investigate the problem of finding the least informative prior on $(\Theta, \mathfrak{B}(\Theta))$ which incorporates that information. We use the standard information theoretic measure of deviation of two probability measures, and write our problem as minimizing the information in the prior subject to a set of constraints.¹

The idea of the paper is similar in spirit to the so called consensus literature (e.g. see survey in Genest and Zidek (1986)). One difference between our approach and the consensus literature is that the latter worries about combining Bayesian priors into an overall prior. To the best of my knowledge, there has not been an attempt in the consensus literature to employ information theoretic measures in combining the priors. Moreover, in the framework of this paper, we do not necessarily have a collection of true Bayesian priors, but rather a collection of (classical or Bayesian) point estimation experiments. If a decision maker were to need estimates of probabilities that the “true” value of the parameter $\theta_0 \in \Theta$ was in some set, he may consult a single estimation experiment, and use it to construct that probability. Given a number of such estimation experiments, our decision maker may want to compute his Bayes risk using his own prior information together with the information in the estimation experiments at hand. We offer one approach to obtain a revised prior on the basis of which to compute his Bayes risk.

¹This is a well accepted Bayesian procedure. Zellner (1988) uses the minimum information formalism to derive and justify Bayes’s inversion formula. For discussions of the merits of that measure of information in the discrete case, see Shannon and Weaver (1962), for more general cases, see Levine and Tribus (1981), Rosenkrantz (1983), and Shore and Johnson (1980) and the references therein.

Another literature that bears similarity to the purpose of this paper is the Meta-Analysis literature (e.g. see Wolf (1986) and the references therein). There, a number of hypothesis testing experiments are combined to reach a decision (reject or fail to reject) based on the p -values of the collection of hypothesis testing experiments. One may think of this paper as an analog of Meta-Analysis for a more general class of decision problems.

We classify the problems requiring a revised prior into three main categories:

1. The decision maker needs a revised prior to be used in a statistical experiment, either for design or updating purposes. In other words, our decision maker is a statistician who needs the entire prior on $(\Theta, \mathfrak{B}(\Theta))$.
2. The decision maker needs to compute a Bayes risk $E_{prior}[r(d, \theta)]$ to choose an optimal decision d . In other words, our decision maker needs to compute the expectation of some collection of functions of θ ($r(d, \theta)$ at a collection of potential decisions $d \in D$).
3. The decision maker wants a point estimate of θ . If the mean of the prior is our point estimate of choice, this reduces to the problem 2. If the mode of the prior is our point estimate, then we need to find the point at which the prior peaks (instead of the whole density as problem 1 requires).

In section 2 of this paper, we rigorously define the problem at hand as a problem of finding the density on Θ that minimizes information subject to a finite collection of constraints, and, in Theorem 1, we find its solution. In section 3, we give three algorithms that will readily resolve the problems for decision makers interested in problems 1-3 above. Finding the whole density for problem 1 turns out to be a calculus of variations problem on a potentially multivariate parameter space Θ . Our Algorithm 1 of section 3 will be shown to yield the desired density. The algorithm starts with an initial guess for the density and converges to the desired minimum information density. The algorithm is ergodic, so if we are interested in problem 2 above, we can simply take the average of $r(\theta, d)$ over the stages of the algorithm to obtain a consistent estimate of the Bayes risk of decision $d \in D$. If we are interested in the mode of the minimum information density, Algorithm 2 offers an alteration of Algorithm 1 that will converge to that mode. In section 4, Theorems 2 and 3, we prove the convergence of those algorithms. Versions of the results in Theorems 1, 2, and 3 exist in the literature in simpler or more complicated contexts. We discuss the relevant literature in which similar results exist, and the necessary adjustments to obtain the results that we need for this paper, prior to the proof of each of the theorems. Section 5 concludes the paper with some cautionary remarks.

2 The Problem

Let our parameter space be $(\Theta, \mathfrak{B}(\Theta))$. Let $\Theta \subset \mathfrak{R}^l$, and let Π_0 be our prior measure on the parameter space. Let Π_0 be absolutely continuous with respect to the Lebesgue measure on \mathfrak{R}^l , and let π_0 be its density (Radon-Nikodym derivative) with respect to that measure. Now, we observe the outcome of a collection of point estimation experiments. We index those experiments by $\alpha = 1, \dots, M$. The information from each experiment summarized by two items:

1. A probability measure Π_α on $\sigma(A_1, \dots, A_n)$. Where A_1, \dots, A_n is a finite partition of Θ , with $A_i \in \mathfrak{B}(\Theta)$, and $\sigma(A_1, \dots, A_n)$ is the smallest sigma algebra generated by the sets A_1, \dots, A_n . Then, clearly, Π_α is fully described by $\Pi_\alpha(A_i)$ for $i=1, \dots, n$.
2. A confidence value p_α .

Example:

If each of the estimation experiments $\alpha \in \{1, \dots, M\}$ is a minimum χ^2 estimation experiment, then the p_α 's can be 1- p -value from the χ^2 distribution of the optimand. Such experiments typically lend themselves to Central Limit Theorems à la LeCam (1986, pp. 305-323), and one can use the asymptotic approximation $\sqrt{T}(\hat{\theta}_\alpha - \alpha_0) \sim N(0, \Sigma_\alpha)$ to obtain the probabilities $\Pi_\alpha(A_i) = \int_{A_i} N(d\theta; 0, \Sigma_\alpha)$, which we would have used for hypothesis testing on the possible position of θ_0 .

We then wish to choose a density $\pi(\theta)$ which is as close as possible in its informative level to our prior density $\pi_0(\theta)$, subject to the constraints that the mass in each of the sets $A_i, i = 1, \dots, n$ is proportional to the amount of mass in that set predicted by our collection of experiments. The predicted mass from each of the experiments $\pi_\alpha(\cdot)$ will be weighted by its appropriate confidence level p_α , and an extra subjective parameter a_α reflecting our qualitative assessment of experiment α . Typically, the a_α 's will all be set to the same value. The entropic measure of information that we use is the standard one, as in Kullback and Liebler (1951), measured by:

$$\mathfrak{I}(\pi, \pi_0) = \int_{\Theta} \pi(\theta) \log\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right) d\theta$$

Our problem, therefore is:

$$\begin{aligned} \min_{\pi} \quad & \int_{\Theta} \pi(\theta) \log\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right) d\theta \\ \text{s.t.} \quad & \int_{\Theta} \pi(\theta) d\theta = 1 \\ \text{and } \Pi(A_i) = \int_{A_i} \pi(\theta) d\theta \propto & \sum_{\alpha=1}^M a_\alpha p_\alpha \Pi_\alpha(A_i) ; i = 1, \dots, n \end{aligned}$$

This is a calculus of variations problem for a function of many variables $\pi(\theta)$, $\Theta \subset \mathbb{R}^l$. Its solution is found in the following theorem.²

Theorem 1

The minimum information density π resulting from solving the above calculus of variations problem is the (Gibbs) density

$$\pi(\theta) = p_U(\theta) = \frac{e^{U(\theta)}}{\int_{\Theta} e^{U(\theta)} d\theta}$$

where $U(\theta) = \log(\pi_0(\theta)) - \sum_{i=1}^n \lambda_i I_i(\theta)$ for appropriate weights λ_α ; and $I_i(\theta)$ is the indicator function for A_i .

Proof:

Write the Lagrangian for our minimization problem in the form

$$\begin{aligned} L(\pi, \theta) &= \int_{\Theta} \left(\pi(\theta) \log(\pi(\theta)) - \pi(\theta) \log(\pi_0(\theta)) + \lambda \pi(\theta) + \sum_{i=1}^n \lambda_i \pi(\theta) I_i(\theta) \right) d\theta - \lambda - \sum_{i=1}^n \lambda_i c_{\alpha} p_{\alpha} \\ &= \int_{\Theta} f(\theta, \pi(\theta), \nabla \pi(\theta)) d\theta + other \end{aligned}$$

where the second term “*other*” contains all the terms that do not depend on $\pi(\theta)$. Since both f and π are C^2 , we can automatically generalize the first variation of standard single integral calculus of variations to get the equations (e.g. Morrey (1966))

$$\sum_{i=1}^n \frac{\partial}{\partial \theta_i} f_{\nabla \pi_i, j} = f_{\pi_j}.$$

Since $\nabla \pi(\theta)$ does not appear in $L(\pi, \theta)$, and since $\pi: \Theta \rightarrow \mathbb{R}_+$ (we obviously do not need to include a non-negativity constraint in the minimization problem since \ln of negative numbers do not exist), the first order conditions for a minimum reduce to $f_{\pi} = 0$. This can simply be written after collecting terms as

$$\log(\pi(\theta)) - \log(\pi_0(\theta)) + (1 + \lambda) + \sum_{i=1}^N \lambda_i I_i(\theta) = 0$$

²In a much simpler context in cognitive science with binary variables, a similar theorem was referred to by Smolensky (1986 in Rumelhart and McClelland ((eds.) 1986)) as the competence theorem. The structure of the problem and the proof methodology for that theorem are simply continuous analogs of the proof in Smolensky (1986 in Rumelhart and McClelland ((eds.) 1986)), although the mathematical results invoked are more advanced. A more general theorem on projection of measures is proven by Csiszár (1975) using more advanced mathematical techniques.

It is obvious that the second order condition for a minimum is satisfied by positivity of π . Now using the first constraint that the density should integrate to unity, and setting $U(\theta) = \log(\pi_0(\theta)) - \sum_{i=1}^n \lambda_i I_i(\theta)$, and $Z^{-1} = e^{-1-\lambda} = 1/[\int_{\Theta} e^{U(\theta)} d\theta]$, we get

$$\pi(\theta) = Z^{-1} e^{U(\theta)}$$

which finishes the proof. ■

Example Continued:

We now construct a very simple example to illustrate the concepts involved. Assume that $\Theta = [-10, 10]$, and let there be three estimation experiments available. The three experiments were minimum χ^2 , and resulted in estimates $\hat{\theta}_1 = 1.0$, $\hat{\theta}_2 = 2.0$ and $\hat{\theta}_3 = 3.0$. The corresponding p_α 's are obtained by 1- p -value of the χ^2 specification test, and turn out to be $p_1 = 0.6$, $p_2 = 0.8$, $p_3 = 0.7$. Let the sample sizes for the three experiments be large enough that we can use the normal approximation, and for simplicity, let $(\hat{\theta}_\alpha - \theta_0) \sim N(0, 1)$. Let the two sets that we care about be $A_1 = [-10, 0)$, and $A_2 = [0, 10]$ (i.e. we only want to make sure that we use information about whether θ is positive or negative). Then, by Theorem 1, we know that our minimum information density subject to the informational constraints is of the form:

$$\pi(\theta) = p_U(\theta) = \pi_0(\theta) e^{-\lambda_1 I_1(\theta)} e^{-\lambda_2 I_2(\theta)}$$

where $I_i(\theta)$ is the indicator function for the set A_i . Further simplify the example by letting $\pi_0(\theta) = 1/20$ be our uninformative prior before observing the results of the point estimation experiments. We can now calculate the $\Pi_\alpha(A_i)$'s easily as follows: $\Pi_1(A_1) = \Phi(-1) = 0.1587 = 1 - \Pi_1(A_2)$, $\Pi_2(A_1) = \Phi(-2) = 0.0228 = 1 - \Pi_2(A_2)$, and $\Pi_3(A_1) = \Phi(-3) = 0.0013 = 1 - \Pi_3(A_2)$. For this simple example, we can directly solve for the λ_i 's from the constraints:

$$\begin{aligned} \Pi(A_1) &= \int_{A_1} \pi(\theta) d\theta \propto e^{-\lambda_1} = \sum_{\alpha=1}^3 p_\alpha \Pi_\alpha(A_1) \\ &= 0.6 \times 0.1587 + 0.8 \times 0.0228 + 0.7 \times 0.0013 = 0.11917 \end{aligned}$$

and

$$\begin{aligned} \Pi(A_2) &= \int_{A_2} \pi(\theta) d\theta \propto e^{-\lambda_2} = \sum_{\alpha=1}^3 p_\alpha \Pi_\alpha(A_2) \\ &= 0.6 \times 0.8413 + 0.8 \times 0.9772 + 0.7 \times 0.9987 = 1.98083 \end{aligned}$$

Normalizing, we get $\Pi(A_1) = 0.11917/2.1 = 0.0567$, $\Pi(A_2) = 1.98093/2.1 = 0.9433$. This immediately yields $\lambda_1 = -\log(0.0567) = 2.86998$, and $\lambda_2 = -\log(0.9433) = 0.05837$. The old prior was $\pi_0(\theta) = 0.05$ for $\theta \in [-10, 10]$, and now after observing the outcomes of the three estimation experiments, it is adjusted to $\pi(\theta) = 0.00567$ for $\theta \in [-10, 0)$, and $\pi(\theta) = 0.09433$ for $\theta \in [0, 10]$. This gives the solution to problem 1 of the introduction. The solution to problem 2 is given by $E_\pi[r(\theta, d)] = 0.00567 \int_{-10}^0 r(\theta, d) d\theta +$

$0.09433 \int_0^{10} r(\theta, d) d\theta$. For problem 3 of the introduction, the mean $E_\pi[\theta] = -0.00567 \times 100/2 + 0.09433 \times 100/2 = 4.433$, and the mode is any number in $[0, 10]$.

It is clear that the example we just illustrated has been manufactured to lend itself to a direct solution. It must be clear to the reader, however, that with more complicated densities, and/or higher dimensional parameter spaces, such a direct computation will not be feasible. In such circumstances, we need to devise algorithms that are sufficiently sophisticated to solve our problems, but simple enough to be easily implementable. We can now rephrase the problems 1-3 of the introduction as follows:

1. Find a simple algorithm to compute the Lagrangian multipliers λ , and $\lambda_i, i = 1, \dots, n$.
2. Find a simple algorithm to compute expectations with respect to the density $\pi(\theta) = p_U(\theta)$.
3. Find a simple algorithm to compute the mode of the density $\pi(\theta) = p_U(\theta)$.

In the following section, we present three algorithms which will solve these three problems. In section 4, we prove the convergence of these algorithms to the quantities we want to compute.

3 Algorithms

The first algorithm is a variant on the famous Metropolis et al. (1953) algorithm used for Monte Carlo integration of multi-dimensional functions. It gives a method for starting with initial draws from any initial density, and converging to a sequence of draws from $\pi(\theta) = p_U(\theta)$. That convergence is the result of part 1 of Theorem 2. Since it will be shown that Algorithm 1 is ergodic, a direct appeal to the Birkhoff ergodic theorem will ensure that taking averages of a function $r(\theta, d)$ over a single sequence of draws from the algorithm will yield a consistent estimate of $E_\pi[r(\theta, d)]$. This is the result of part 2 of Theorem 2.

Algorithm 1 (Gibbs Sampler)

1. *Start with an initial density f_0 on Θ , and draw θ_0 at random from the density f_0 , this is set to be the 0^{th} iteration.*
2. *At random choose one of the coordinates of θ (i.e. $1, 2, \dots$ or l). At the t^{th} iteration, let that draw be $d_t \in S = \{1, 2, \dots, l\}$.*

3. Update the density f_t at the t^{th} iteration using the following Chapman- Kolmogorov type equation

$$f_{t+1}(\theta_s = x_s; s \in S) = p_U(\theta_{d_{t+1}} = x_{d_{t+1}} | \theta_s = x_s; s \neq d_{t+1}) \cdot f_t(\theta_s = x_s; s \neq d_{t+1})$$

where p_U is the Gibbs density defined by $U(\theta) = \sum_{i=1}^n \lambda_i I_i(\theta) - \log(\pi_0(\theta))$ for appropriate weights λ_i to be discussed later. The conditional density in the updating rule is defined by

$$p_U(\theta_s | \theta_d; d \neq s) = \frac{p_U(\theta)}{\int_{x_s \in \Theta_s} p_U(\theta_1, \dots, \theta_{s-1}, x_s, \dots, \theta_t) \cdot dx_s}$$

where $s, d \in S = \{1, \dots, l\}$ and $\theta \in \Theta = \times_{s=1}^l \Theta_s$.

4. Randomly draw r_{t+1} from the density f_{t+1} , and iterate steps 2 through 4 until convergence of f_t .

In order to get the mode of $\pi(\theta) = p_U(\theta)$ for application 3 of the introduction, we need to solve a problem of maximizing a function of many variables. Since we already have the Gibbs sampler algorithm handy, we have all the machinery for implementing a simulated annealing maximization algorithm. Simulated annealing is a combination of a randomized and a deterministic maximization routines, which, when the randomized component vanishes at the appropriate speed, converges to the global maximum with probability 1. A version of simulated annealing that readily assimilates with our Algorithm 1 (this algorithm is an adaptation of the one in Geman and Geman (1984)) is presented below, and its convergence to the mode of $\pi(\theta)$ is the result of part 3 of Theorem 2.

Algorithm 2 (Simulated annealing)

Implement Algorithm 1, using the Gibbs density $p_{U_t}(\theta)$, where $U_t(\theta) = U(\theta)/T_t$. As the iteration index $t \uparrow \infty$, let $T_t \downarrow 0$ at a sufficiently slow rate.

The remaining problem now is to find an algorithm which will give us the Lagrange multipliers λ and $\lambda_i, i = 1, \dots, n$ which define the function $U(\theta)$. The following algorithm is based on a similar algorithm in Smolensky(1986, in Rumelhart and McClelland ((eds.) 1986)). The idea of the algorithm is to take random draws from p_U with the current values of the λ_i 's, and compare the resulting probabilities of the A_i sets to the given probabilities $\Pi(A_i)$. The λ 's are then adjusted depending on the resulting discrepancy. The convergence of this algorithm is the result of Theorem 3.

Algorithm 3 (Learning)

1. Set all $\lambda_i = 0; i = 1, \dots, n$.

2. Draw an i.i.d. sample ϕ_1, \dots, ϕ_ν from the Gibbs density p_U defined by the current values of the λ_i 's.
3. Compute a decrement $d_i = \frac{1}{\nu} \sum_{j=1}^{\nu} I_i(\phi_j)$.
4. Update $\lambda_i \leftarrow \lambda_i + \sum_{\alpha=1}^M c_{\alpha} p_{\alpha} \Pi_{\alpha}(A_i) - d_i$; $i = 1, \dots, n$.
5. Repeat steps 2-4 until convergence of the λ_i 's.

4 Convergence of the algorithms

The first part of Theorem 2 proves convergence of Algorithm 1 to the minimum information density $p_U(\theta)$ when the λ_i 's are given. The second part proves the convergence of averages of any integrable function $r(\theta, d)$ over the sequence of draws in Algorithm 1 to its expectation $E_{\pi}[r(\theta, d)]$. The third part proves the convergence of Algorithm 2 to the mode of the minimum information density $\pi(\theta) = p_U(\theta)$. The third part of the theorem follows Geman and Geman (1984), and provides the convergence to the mode at a very slow speed of $T_t \downarrow 0$, and the simulations in their paper suggest that a much faster rate of convergence should be possible, but such results are not yet available. For the purposes of part 3 of the theorem, let us define

$$R_0 = \{\theta \in \Theta: U(\theta) = \max_{\phi \in \Theta} U(\phi)\}$$

and let $\bar{\pi}$ be the uniform measure on R_0 . Also define $U^* = \max_{\theta \in \Theta} U(\theta)$, $U_* = \min_{\theta \in \Theta} U(\theta)$, and $\Delta = U^* - U_*$.

Theorem 2

1. Algorithm 1 implemented with transition p_U (of step 3 in the algorithm, with U defined by the λ_i 's of theorem 1) converges, for all initial densities f_0 to π .
2. Let g be any L^1 function $\Theta \mapsto R$, then $\frac{1}{t} \sum_{\tau=1}^t g(\theta_{\tau})$ converges almost everywhere to $E_{\pi}[g(\theta)]$.
3. In Algorithm 2, let $T_t \rightarrow 0$ and $T_t \geq N \cdot \frac{\Delta}{\ln(t)}$ for all $t \geq t_0 \geq 2$. Then, for all initial densities f_0 ,

$$\lim_{t \uparrow \infty} f_t(\theta_t = x | \theta_0 = y) = \bar{\pi}(x)$$

Proof:

It is clear that the stochastic process θ_t is a Markov process. Now, for a fixed t , let $\theta \in \Theta = [0, 1]^n$, and for any $x \in [0, 1]$; let θ^x be the vector such that $\theta_{d_t}^x = x$, and $\theta_s^x = \theta_s$

for all $s \neq d_t$. Then, the transition kernel for the Markov process θ_t is defined for all $\theta' \in \theta^x$

$$K(\theta'|\theta) = P_{d_t}.p_U(\theta'_{d_t} = x_{d_t}|\theta'_s = \theta_s; s \neq d_t)$$

where $p_U(\cdot)$ and $p_U(\cdot|\cdot)$ are defined as before, and where P_{d_t} is the probability of choosing coordinate d_t to change at time t (this will typically be $1/l$). But that kernel easily shows us that

$$p_U(\theta').K(\theta|\theta') = p_U(\theta).K(\theta'|\theta)$$

where θ and θ' differ in at most one component, and K is a stationary kernel. Hence, if we start with the density $f_t(\cdot) = p_U(\cdot)$, we get

$$\begin{aligned} f_{t+1}(\theta) &= \int_{\theta' \in \theta^x} K(\theta|\theta').p_U(\theta').d\theta' \\ &= \int_{\theta' \in \theta^x} K(\theta'|\theta).p_U(\theta).d\theta' = p_U(\theta) = f_t(\theta) \end{aligned}$$

and hence the Gibbs density $p_U(\cdot)$ is a stationary density. Notice moreover, that in Algorithm 1, all coordinates will be visited infinitely often with probability 1, and by the positivity of $p_U(\cdot)$ and hence of $p_U(\cdot|\cdot)$, $K(\cdot|\cdot)$ is irreducible, and p_U defines an irreducibility measure (by Nummelin (1984, proposition 2.4, p.13)). By the positivity of p_U , and since we only consider densities of non-atomic measures, the measure defined by p_U is a maximal irreducibility measure (i.e. all other irreducibility measures are absolutely continuous with respect to it). We also know that all sets A with $\int_A p_U(\theta)d\theta > 0$ will be visited infinitely often with probability 1. But that is equivalent to positive Harris recurrence (see Nummelin (1984, Ch. 3)), and that together with the aperiodicity guaranteed by $p_U > 0$ implies the Harris ergodicity of the process θ_t . Since we know that p_U is a stationary density for $K(\cdot|\cdot)$, it follows that p_U is the unique stationary density and the limiting density of the process (by Harris (1956, Theorem 1)). This concludes the proof of part 1. Part 2 follows immediately from the Harris ergodicity of the heat bath, and the Birkhoff ergodic theorem (e.g. Walters (1982, pp.34-35)).

For part 3, the proof of theorem B in Geman and Geman (1984) extends to the continuous state space without requiring any changes other than replacing all transition probabilities with transition kernel densities, and replacing all sums with integrals where applicable. ■

Theorem 2's results were proven when the θ_t 's are drawn from the density $p_U(\theta)$ defined by the correct λ_i 's. To compute the appropriate λ_i 's, Algorithm 3 draws a sequence of ϕ_t from Algorithm 1 with any other values for the λ_i 's. Following the same proof of Theorem 2, this sequence can be used to compute consistent estimates of $\Pi(A_i)$'s under the density with the (potentially wrong) λ_i 's. We can then adjust the λ_i 's using a Newton's method algorithm, and check their predicted $\Pi(A_i)$'s again, and so on, until we get the appropriate values of the λ_i 's. This is the procedure detailed in Algorithm 3, and the following Theorem proves its convergence to the correct values of the λ_i 's.³

³The proof of this result is similar to that for the discrete case provided in Smolensky (1986, in: Rumelhart and McClelland ((eds.) 1986)).

Theorem 3

Algorithm 3 converges as $\nu \uparrow \infty$, and the number of iterations $t \uparrow \infty$ to the λ_i 's of Theorem 1.

Proof:

It is clear that the λ_i 's of Theorem 1 are those that minimize the convex function

$$F(\lambda_1, \dots, \lambda_n) = \log \int_{\Theta} e^{\left(\sum_{i=1}^n \lambda_i [\Pi(A_i) - \sum_{\alpha=1}^M c_{\alpha} p_{\alpha} \Pi_{\alpha}(A_i)] \right)} d\theta$$

By differentiating under the integral, the first order conditions for λ_i 's imply $p_U = \pi$. The second order conditions for a minimum are guaranteed by convexity of F (by the positivity everywhere of p_U). Then steepest descent to that vector of λ_{α} 's which minimizes F is simply

$$\frac{d\lambda_i}{dt} \propto -\frac{\partial F}{\partial \lambda_i} = \sum_{\alpha=1}^M c_{\alpha} p_{\alpha} \Pi_{\alpha}(A_i) - \int_{\Theta} I_i(\theta) p_U(d\theta)$$

As $\nu \uparrow \infty$, sample means computed for the decrement of Algorithm 3 converge to the second expectation term of the last expression by part 2 of Theorem 2. As $t \uparrow \infty$, convergence follows by the convexity of F . \blacksquare

5 Concluding Remarks

In the exposition of the general framework in section 2, we abstracted away from a number of problems that have to be taken into consideration when empirically applying the suggested methods. The main criticisms that apply to meta-analysis also apply here. We quote the following major criticisms as grouped by Glass et al. (1981) and stated in Wolf (1986, p.14):

1. *Logical conclusions cannot be drawn by comparing and aggregating studies that include different measuring techniques, definitions of variables ... because they are too dissimilar.*
2. *Results of meta-analyses are uninterpretable because results from "poorly" designed studies are included along with results from "good" studies.*
3. *Published research is biased in favor of significant findings because nonsignificant findings are rarely published ...*

To answer these criticisms, let us note:

1. In our construction of the framework of combining a number of estimators, one can argue that despite the fact that the results may be incomparable, a comparison is necessary and is indeed done in a non-statistical way. Even though the resulting beliefs about the parameter estimates θ_α 's are not Bayesian posteriors, they are being treated as such by practitioners. Hypothesis testing, ...etc. is performed as though the asymptotic distribution of an estimate $\hat{\theta}$ about θ can be reversed to do inference about the probability that the "true" θ lies in any particular region. Given that such practices are followed in a special case of our technique where one particular study is given the full probabilistic weight, we suggest that our procedure should be useful by giving us the flexibility to assign different sets of weights to the different studies.
2. The flexibility of assigning the relative weights a_α to the various studies allows us to give less weight to poorly designed studies (if we indeed could determine that they were poorly designed). This makes the second criticism much less destructive.
3. If it is true that published research is biased in favor of results with a high p_α , then that should actually help us by reducing the number of studies whose weights $ca_\alpha p_\alpha$ *ceteris paribus* (i.e. keeping the quality of the study and hence the relative weight a_α constant) are low, and hence whose contribution to our posterior π should be minimal. Other types of biases in favor of certain hypotheses admittedly still pose a problem.

References

- Csiszár, I. 1975. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability* 3, no.1:146–158.
- Geman, S. and D. Geman. 1984. Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.
- Genest, C. and J. Zidek. 1986. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science* 1:114–148.
- Glass, G., B. McGraw, and M. Smith. 1981. *Meta analysis in social research*. Beverly Hills: Sage Pub.
- Harris, T. 1956. The existence of stationary measures for certain Markov processes. University of California Press, Berkeley.
- Kullback, S. and R. Liebler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22:79–86.
- LeCam, L. 1986. *Asymptotic methods in statistical decision theory*. New York: Springer-Verlag.
- Levine, R. and M. Tribus. 1981. *The maximum entropy formalism*. Cambridge, MA: M.I.T. Press.
- Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* 21.
- Morrey, C. 1966. *Multiple integrals in the calculus of variations*. New York: Springer-Verlag.
- Nummelin, E. 1984. *General irreducible markov chains and non-negative operators*. Cambridge: Cambridge University Press.
- Rosenkrantz, R. 1983. *E. T. Jaynes: Papers of probability, physics, and statistical physics*. Dordrecht: Reidel Pub. Co.
- Rumelhart, D. and J. McClelland. (eds.) 1986. *Parallel distributed processing vols. I and II*. Cambridge, MA: M.I.T. Press.
- Shannon, C. and W. Weave. 1962. *The mathematical theory of communications*. Urbana: University of Illinois Press.
- Shore, J. and R. Johnson. 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross entropy. *IEEE Transactions on Information Theory* IT-26:26–37.
- Walters, P. 1982. *An introduction to ergodic theory*. New York: Springer-Verlag.

- Wolf, F. 1986. *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills: Sage Pub.
- Zellner, A. 1988. Optimal information processing and Bayes's theorem. *The American Statistician* 42:278–284.